Programación Python para Big Data

Lección 6: VAEX y DASK

eip ESCUELA INTERNACIONAL POSGRADOS

Programación Python para Big Data

ACTIVIDAD LECCIÓN 6

Objetivos

- En el presente tema hemos hablado de algunos conceptos básicos de VAEX y de DASK, que son 2 muy buenas alternativas a PANDAS cuando necesitamos manejar un set de datos con muchas filas.

- El objetivo principal será aprender bastante bien estas herramientas mencionadas: VAEX y DASK

- Otro objetivo será fomentar la habilidad proactiva de el/la alumno/a quien tendrá la oportunidad de aprender más cosas y con ello obtener mejor nota.

Contenido correspondiente a Lección 6:

1. Herramientas para manejar grandes volúmenes de datos
   - **1.1.** VAEX
   - **1.2.** DASK
   - **1.3.** Comparativas usando: VAEX, DASK y PANDAS

Actividad relacionada con la Lección 6:

PUNTUACIÓN MÁXIMA QUE SE PUEDE OBTENER: 12 PUNTOS

Obviamente la máxima calificación será un 10.

Primera parte de la Actividad (Hasta 6 puntos)

El/la alumno/a deberá enviar un **archivo .ipynb** con EL MISMO EJERCICIO hecho en el manual, indicando el tipo de PC con el que se ha hecho el experimento. (Obviamente no es necesaria la comparativa que se ha presentado en el manual haciendo alusión al tipo de Disco Duro).

Usa el Disco Duro y Sistema Operativo que quieras, pero detalla qué has empleado de la misma forma que lo has visto en el manual.

Por ello se hará con PANDAS, con VAEX y con DASK

En el mismo .ipynb si lo deseas, puedes dejar alguna celda en blanco, y continuar..

Segunda parte de la actividad (Hasta 2 puntos)

A elegir entre VAEX y DASK se puede entregar lo que se quiera, tratando de aprender un poco más de lo ya visto.

Es posible buscar si se quiere otro set de datos, para ello.

El objetivo no es copiar y pegar cualquier cosa y entregar la actividad para evaluación, sino tratar de aprender alguna cosa más de forma autodidacta.

De modo que hay varias opciones, por ejemplo:

- hacer lo mismo con otro set de datos de gran volumen. (Pudiera ser algo más pequeño que éste si no se encuentra uno mayor)
- Buscar más cosas de VAEX y DASK no mencionadas en el manual y emplearlo. (Gráficas, tiempos para “groupby”, o lo que se quiera).
- Etc.

En el mismo .ipynb si lo deseas, puedes dejar alguna celda en blanco, y continuar..

Tercera parte de la actividad (Hasta 2 puntos)

Modin es una gran herramienta que existe para determinados sets de datos.

Se pide que se haga el mismo ejercicio del manual con esta herramienta.

NOTA: Solo es necesario mostrar el tiempo de lectura del .csv con modin, de 2 formas diferentes. (Cuando leas la documentación entenderás rápidamente a qué nos referimos).

En el mismo .ipynb si lo deseas, puedes dejar alguna celda en blanco, y continuar..

Cuarta parte de la actividad (Hasta 2 puntos)

A elegir entre Rapids ó Koalas (aunque pudiera ser otro distinto a los mencionados si se prefiere).

Revisar en internet su documentación, y elegir uno de ellos, del cual se va a hacer algo (lo que se quiera) en base a lo que el/la estudiante aprenda.

Notas:

- Obviamente, no pueden ser solo 2 líneas, aunque tampoco se pide que lleve excesivo tiempo. Lo que se busca es ver que disponemos de varias alternativas, y, cuando aprendamos varias, nos especializaremos en la que más nos guste.
- Es posible añadir texto, (y no solamente código), de cosas que llamen la atención, de cosas que sean importantes o a tener en cuenta, etc.